An Update of the Manually Annotated Amharic Corpus

Pavel Rychlý Gezahegn Tsegaye Lemma

Natural Language Processing Centre
Faculty of Informatics, Masaryk University
University of Calabria

December 8, 2018

Walta Info Corpus (WIC)

- Amharic language
- about 210,000 words in 1,065 documents
- texts from the Web news published by the Walta Information Center in 2001
- two versions in two scripts (Fidel, Sera)
- manually annotated
  complicated process using pen and paper, retyping

HaBiT project

- corpora and tools for less resourced languages
- 4 Ethiopian languages, Czech, Norwegian
- Walta Info Corpus from Oslo
- unified two versions into one
- cleaning
- release at Lindat/Clarin
- accessible at HaBiT installation of SkE

WIC PoS tag-set

- part of speech only
- 30 different tags
- prepositions and/or conjunctions are attached to words in Amharic
- PoS tags annotate prepositions and/or conjunctions
- nouns (tag N)
  - noun with a preposition as prefix (tag NP)
  - noun with a conjunction as suffix (tag NC)
  - noun with a preposition as prefix and a conjunction as suffix (tag NPC)

PoS tag ambiguity in WIC

- 34,000 types
- 20,000 hapax legomena
- 4600 types with at least two different PoS tags

በተለይም (*specifically*)		ያህል (*about*)	
ADJ	5	ADJ	5
ADJC	1	ADV	15
ADJP	3	N	11
ADV	141	NC	3
CONJ	3	NP	1
N	1	PREP	2
NC	4	PRON	1
NP	24	UNC	37
NPC	4	V	48
UNC	25	VP	2
VPC	1	VREL	2

Error correction

- 68 highly ambiguous words
- 200 combinations of word-tag
- 5000 tokens
- 139 word-tag combination incorrect
- 2,300 tokens corrected

Error correction

- 68 highly ambiguous words
- 200 combinations of word-tag
- 5000 tokens
- 139 word-tag combination incorrect
- 2,300 tokens corrected
- ሁለት (*two*): 139/150 correct hits, 11 incorrect (3 different tags)
- ለመስራት (*to work, to make*) 34/40 incorrect hits

Evaluation

- 10-fold cross validation
- higher accuracy
- TreeTagger: from 87.4 to 87.9
- APtagger: from 82.9 to 83.6

Conclusion

- clean data are better data
- even small change in data (2,300 tokens from 200,000) could mean a significant improvement in tagging

Bonus: APtagger

- Averaged Perceptron tagger
- very simple, 200 lines of Python code